Open Praxis, vol. 6 issue 2, April–June 2014, pp. 159–170 (ISSN 2304-070X)
2014 OCWC Global Conference Selected Papers

Crowd-sourcing (semantically) Structured Multilingual Educational Content (CoSMEC)

Darya Tarasowa & Sören Auer
University of Bonn (Germany)
darya.tarasowa@gmail.com & auer@cs.uni-bonn.de

Ali Khalili & Jörg Unbehauen
University of Leipzig (Germany)
khalili@informatik.uni-leipzig.de & unbehauen@informatik.uni-leipzig.de

Abstract

The support of multilingual content becomes crucial for educational platforms due to the benefits it offers. In this paper we propose a concept that allows content authors to use the power of the crowd to create (semantically) structured multilingual educational content out of their material. To enable the collaboration of the crowd, we expand our previously developed CrowdLearn concept and WikiApp data model to support multilingual educational content. The expanded concept, CoSMEC, was evaluated with an example implementation within the web-based educational platform SlideWiki. Based on this experience, we provide solutions for the most complicated technical issues we faced during the implementation. This paper also discusses statistics which show the flow of multilingual content usage.

Keywords: OER; multilinguality; collaborative authoring

Introduction

Nowadays, the support of multilingual content becomes crucial for educational platforms because of the benefits it offers. One of the most important benefits of multilingual content supplying is an opportunity to educate without language barriers. This causes the number of learners to increase according to the number of languages content is available in. The learning quality for previously presented users increases as well, due to the possibility of studying in their mother language. Another benefit of the multilinguality is the ability to share and exchange knowledge within a multicultural environment. This opportunity is crucial for fostering the increase of education quality, especially in developing countries. Institutes and universities in these areas often base courses on outdated information, concepts and theories. Access to the international scientific knowledge through content synchronization would help experts and educational institutes to be aware of the state-of-art in their field.

The creation of high-quality educational content is a time and resource consuming task. The task requires even more resources if there is a need to offer the content in different languages. A winning strategy in this case is to produce the content collaboratively. The collaboration can be established within a limited group of experts (1) or involve the power of a crowd (2). The first approach has a number of issues:

- it requires a lot of preparation, as the domain experts with the knowledge of certain languages have to be found first;
- each expert has her own background and point of view on the topic, thus, the negotiations can have a negative effect on the time costs;

Reception date: 25 February 2014 • Acceptance date: 8 April 2014
DOI: http://dx.doi.org/10.5944/openpraxis.6.2.117

- it results in production of content available in only a limited number of predefined languages;
- the content easily becomes outdated, as a limited number of experts being spread internationally might not be aware of all new findings in the domain;
- the content refinement is time-consuming and can cause de-synchronization of the content.

To overcome these issues, we propose to use the power of a crowd to author the content available in a number of different languages. In this case the list of available languages is not fixed and the number of languages depends on the community size, diversity and activity. An example of the successful application of crowd-sourcing techniques to multilingual content authoring is Wikipedia (http://wikipedia.org). However, the Wikipedia concept does not suppose an (semi-)automatic synchronization of the content presented in different languages. Although synchronization is sometimes being done manually by contributors, the versions in different languages often stay contradictory to each other.

In order to facilitate content synchronization, we propose the content to be semantically structured. By semantic structuring we mean the splitting of the content into elements in such a way, that each of them fully covers an individual piece of knowledge. To organize the content in this way we propose to use our previously developed CrowdLearn concept (Tarasowa *et al.*, 2013). CrowdLearn enables collaborative authoring of semantically structured educational content. The concept is an application of crowd-sourcing techniques to e-learning content creation, re-purpose and reuse. It includes the WikiApp data model to support the versioning of structured content objects. Furthermore, CrowdLearn supports social networking activities and enables users to proactively interact with each other to acquire knowledge.

In the current paper we expand the CrowdLearn concept to support multilingual educational content. The expansion requires the defining of additional concepts and relations for the WikiApp data model. We named the expanded version of the CrowdLearn concept 'CoSMEC—Crowd-sourcing (semantically) Structured Multilingual Educational Content' to reflect the support of multilinguality. The paper is structured as follows: we first describe our CoSMEC concept. Then, we present an example implementation of the concept and describe our solutions for the most complicated implementation issues. We evaluate the concept and discuss the results in "Evaluation" section. After that we show the connections with the current research in the field. In the end, we summarize and conclude our work.

Concept

The proposed CoSMEC concept combines and expands several existing paradigms. The *paradigm of Open Educational Resources* supposes the creation and sharing of reusable, re-purposable learning objects, annotated by standardized metadata. The *wiki paradigm* allows every user to refine the published content by creating a new version after every edit. The *crowd-sourcing technique* supposes version control of the content and allows branching and merging of the revisions. In order to efficiently apply the paradigms above, the concept requires the content to be highly-structured. CoSMEC enables the support of multilingual content for this complex data model. The overview of the concept is presented in Figure 1.

Collaborative authoring of structured educational content

To enable collaborative authoring of semantically structured educational content, we use our CrowdLearn concept, as described in detail in Tarasowa *et al.* (2013). The CrowdLearn concept is an application of crowd-sourcing techniques to e-learning content creation, re-purpose and reuse. In this paragraph we give a short overview of the concept.

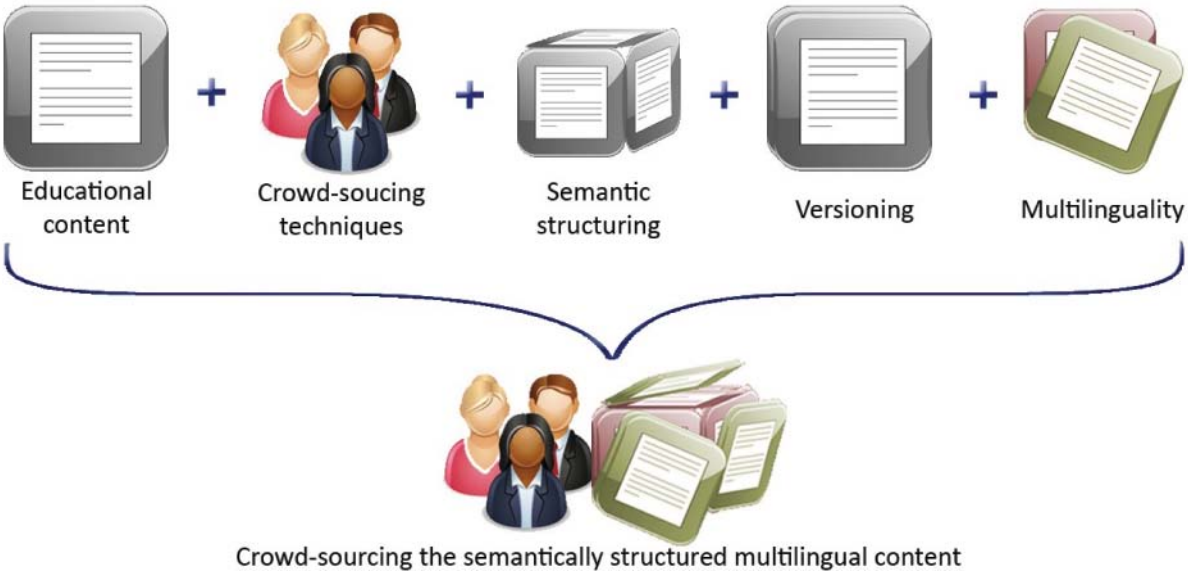

Figure 1: Overview of the CoSMEC concept

As its main innovation, CrowdLearn combines crowd-sourcing techniques with the creation of highly-structured e-learning content. E-learning material, when combined with crowd-sourcing and collaborative social approaches, can help to cultivate innovation by collecting and expressing individual's (contradictory) ideas. In accordance with the CrowdLearn concept, instead of dealing with large learning objects (often whole presentations or tests), we decompose them into fine-grained learning artifacts. Thus, rather than a large presentation, users will be able to edit, discuss and reuse individual slides; instead of a whole text she/he will be able to work on the level of individual questions. This structure efficiently facilitates the reuse and re-purpose of the learning objects.

The concept of decomposition meets the requirements of modern e-learning content standards. This allows the CrowdLearn concept to produce standard adhering content. In particular, the concept adopts the SCORM standard (ADL, 2011a) and practical recommendations (ADL, 2011b) and expands the standard for the collaborative model.

To enable versioning of the content, the CrowdLearn concept proposes the WikiApp data model. The WikiApp is a refinement of the traditional entity-relationship data model. It adds some additional formalisms in order to make *user*s as well as *ownership*, *part-of* and *based-on* relationships first-class citizens of the data model. A set of content objects connected by part-of relations can be arranged and manipulated in exactly the same manner as an individual non-structured object. The model natively supports versioning and structuring of the content objects.

Supporting social networking activities in CrowdLearn enables students to proactively interact with each other to acquire knowledge. Besides increasing the learning process quality, social activities improve the quality of the created learning material. Even when answering a quiz, users can contribute by analyzing the quality of the questions and making suggestions of how to improve them. Thus, knowledge is being created not only explicitly by contributors, but also implicitly through discussions and answering the questions of assessment tests; in other words through native learning activities. Furthermore, social activities enable social network analysis of users (both teachers and learners) and learning objects (Linta *et al.*, 2011). We illustrate the WikiApp data model in Figure 2.

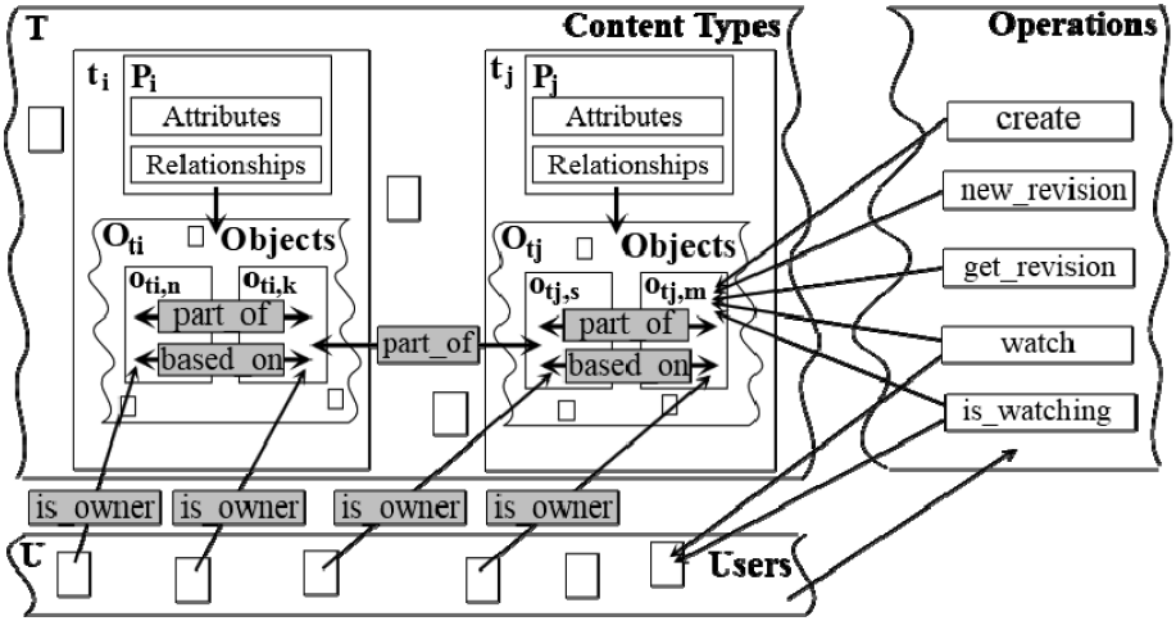

Figure 2: Conceptual view of the WikiApp data model

Managing multilingual educational content

In order to enable the support of multilingual content, CoSMEC allows the content versions to be presented in different languages by adding the *translatedInto* and its inverse *translatedFrom* relations to the WikiApp data model. CoSMEC does not deal with the objects being originally created in different languages. Instead, our concept requires the presence of the source content object, as well as its translations. Enforcing this requirement allows us to:

- distinguish between users' authoring and translating contributions to the content,
- present the list of original content authors in all translations,
- propagate changes in order to synchronize the content of translations with the source.

The CoSMEC concept introduces the paradigm of co-evolution of multilingual content, that means the ability to update a translation to the current state of the source object and vice versa. However, the requirement above only allows users to update the content of a translation according to an original version, but not contrariwise. To overcome this limitation, we enable the translation of an object back to the original language, thus creating a revision of the source object. This back-translation requires a mechanism of merging the revisions, as we do not want to overwrite the original content with a repeatedly-translated one. Thus, in order to function efficiently, the mechanism of co-evolution requires three operations to be defined:

1. initial translation (see Figure 3, box 1);
2. synchronization of the translation with the source (see Figure 3, box 2);
3. merging the revisions (see Figure 3, box 3)

Details of the mechanism implementation are presented in the next section.

Implementation

We implement and evaluate the CoSMEC concept with SlideWiki (http://slidewiki.org)—a web-based crowd-learning platform.

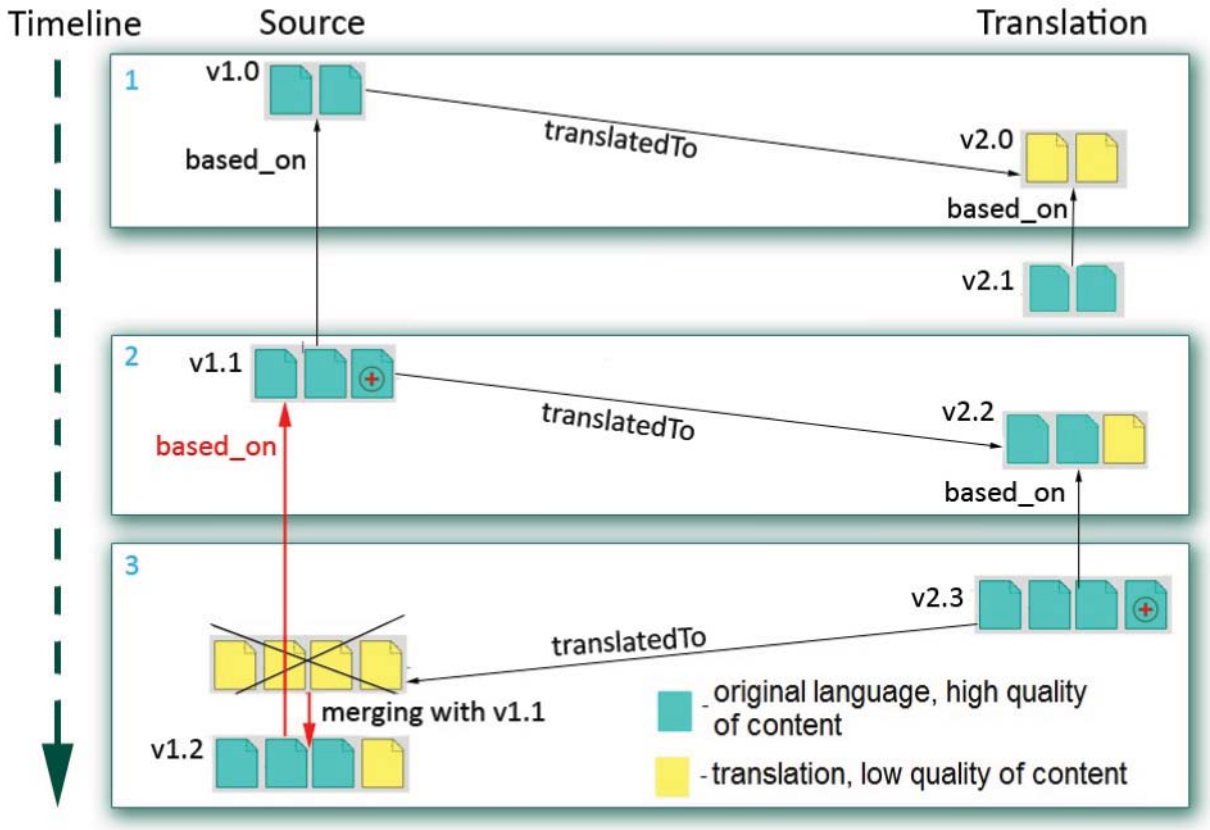

Figure 3: CoSMEC mechanism of the multilingual content co-evolution: 1- initial translation, 2- partial translation to synchronize the content of translation with the source, 3- merging the revisions to avoid the repeatedly translated content

SlideWiki

SlideWiki deals with two types of (semi-)structured learning objects: slide presentations and assessment tests. The content is organized according to the CrowdLearn concept. Thus, SlideWiki uses two implementations of the WikiApp data model. The first implementation is used for managing slides and presentations. It includes individual slides (consisting mainly of HTML snippets, SVG images and meta-data), decks (being ordered sequences of slides and sub-decks), themes (which are associated as default styles with decks and users) and media assets (which are used within slides). The second implementation was developed for managing questions and assessment tests. It includes questions for the slide material (the question is assigned to all slide revisions), tests (which could be organized manually by user or created automatically in accordance with the deck content), and answers (which are part of the questions).

We implicitly connected these two WikiApp instances by adding two relations. Firstly, we assigned questions to slides. Thus, during the learning process users are able to answer the tests and have a look at the assigned slide if necessary. The important issue here is that we assign a question not to an individual slide revision, but to the slide itself. This decision gives an opportunity to create a new slide revision, that already has a list of questions, collected from other revisions. Secondly, we assigned assessment tests to concrete deck revisions. Thus the automatically created test saves the structure of the corresponding deck revision. This allows us to use module-based assessment to score the test results.

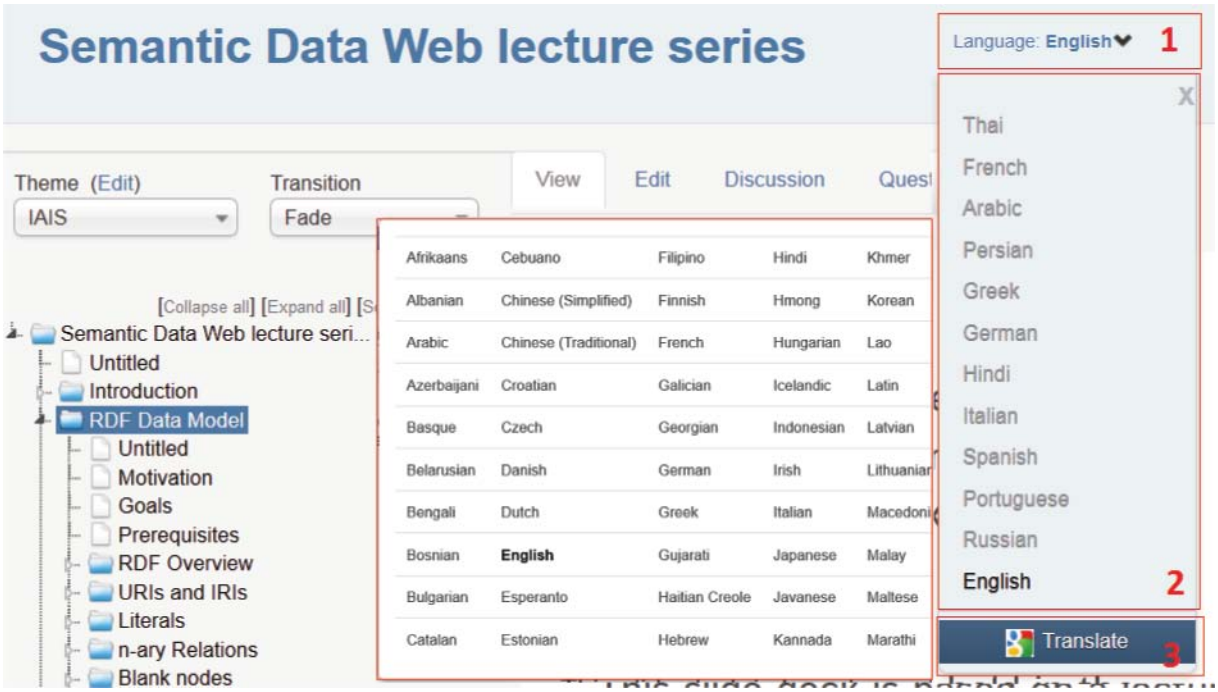

Figure 4: Interface of translations management: 1- language of selected object; 2- a drop-down list with links to languages available for the selected object; 3- button for translation; 4- dynamically updated list of the languages supported by Google Translate service

The basic functionality of SlideWiki is described in details in Tarasowa *et al.* (2013). In the current paper we focus on recently implemented support of multilinguality.

Co-evolution of the SlideWiki content

As was already discussed, the implementation of co-evolution of source object content and its translations supposes the implementation of three operations: initial translation, synchronization and merging of the revisions. In this subsection we describe an example implementation of these operations in SlideWiki.

Translation. Our architecture allowed us to implement a translation operation backed by the Google Translate service (http://translate.google.com). After translation into one of 71 currently supported languages, the presentation can be edited, re-structured and reused independently from its source. The implemented interface of the translation management is presented in Figure 4.

Synchronization. To enable synchronization of original and translated versions, every further revision of translated objects inherits the link to the source revision (see v2.1, Figure 5). The changes in the original version of the object cause the creation of new revision v1.1. Additionally, users are notified of translations that have become out of sync with the source (exclamation marks in v2.0 and v2.1, Figure 5.

In SlideWiki implementation of the synchronization is slightly different for slides and decks. We decided not to implement manual synchronization of decks, as we considered the process to be too complicated for users. Users can only trigger an automatic synchronization. However, our data model allows us to get all existing translations for all existing slide revisions. Then, during an automatic deck synchronization we do not repeatedly translate the slides which have already been

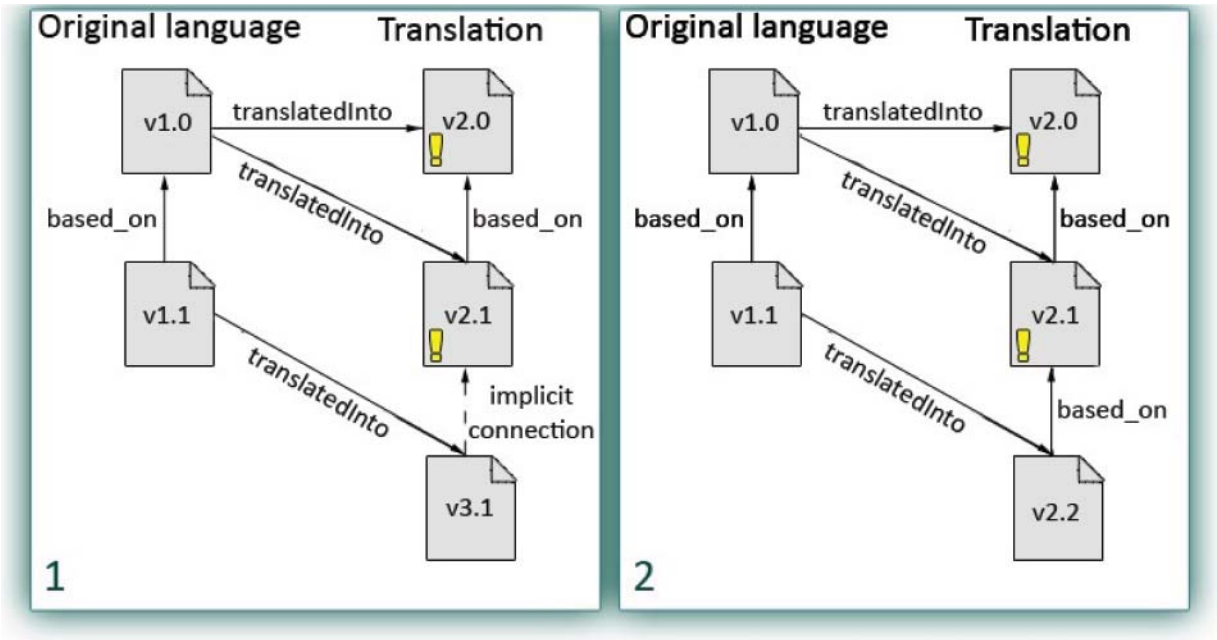

Figure 5: Two scenarios of content synchronization between translation and source: 1- automatic; 2- manual synchronization

translated to the target language before. Instead, we include in the translation the latest revision of the slide in the target language. Thus, the synchronization of decks only adds the slides and subdecks that were not included in the source deck at the moment of previous translation.

For the slides both automatic and manual synchronization are enabled. The users are able to compare v1.0 and v1.1 and decide, either they want to redo the translation (scenario 1 at Figure 5) or they want to update the content manually (scenario 2).

Merging the revisions. SlideWiki implements the revision control in accordance with the WikiApp data model. However, we defined rules and restrictions to increase the performance. We wanted to avoid an uncontrolled proliferation of deck revisions. However, this would happen due to the fact that every change in a slide would also trigger the creation of a new deck revision for all the decks that slide is a part of. In addition, when creating a new deck revision, we always need to recursively spread the change into the parent decks and create new revisions for them if necessary. To deal with this issue, we introduced the content owner and member of editor group roles. If the changes are made by a user belonging to one of these two roles, the creation of a new deck revision is not triggered (the new slide revision however is created).

As we allow the owner of a deck revision to change it without the creation of a new revision, it was an important issue whether we should allow the multiple translation of the same revision into the same language or not. We decided to allow it, however, this led to the situation that we would get several identical presentations with content of bad quality, since it was translated automatically and not edited manually. However, we could not disable the multiple translations, because in that case it would be impossible for example to get translations of new slides if they were added by the owner. Thus, merging the revisions became the crucial operation, not only for merging back-translation with the source, but also for merging multiple translations in the same language. Figure 6 illustrates the interface for comparing and merging deck revisions.

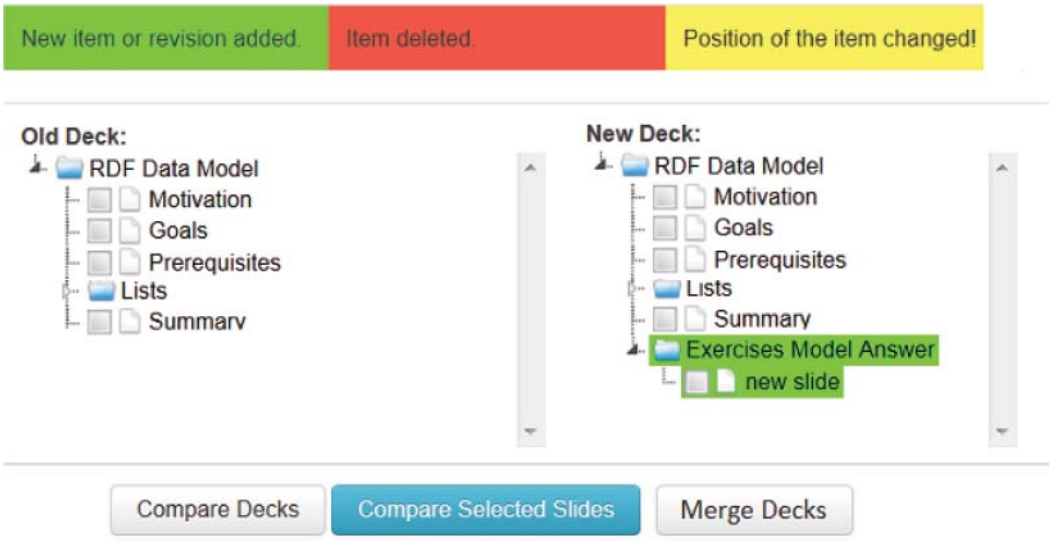

Figure 6: Interface for comparing and merging deck revisions

Evaluation

To evaluate our work, we collected statistics on the usage of the translation tool as well as statistics on the multilingual content it produced.

Figure 7 presents the distribution between original and translated versions in relation to the total number of content objects. As the WikiApp data model does not enable deletion or update of the content, the graphs can be viewed as time trends. The blue line shows an example moment in time, when the SlideWiki database consisted of 16321 slides overall. The graphic shows that 78% of the slides at the moment were in their original language, 22% were translations and 5% of the total number of slides were revised after translation. At the same moment, about 35% of the decks were translations. Thus, the percentage of translated decks increases faster than that of translated slides. This means that the presentations consisting of less than average number of slides are being translated more often. This can be due to the fact that users want to try the feature before using it on large decks. However, the assumption needs further investigation.

According to the statistics, the percentage of content created by translation has a strongly increasing trend. We predict that percentage of translations will soon prevail over the percentage of source objects. From one perspective, the prevalence means decreasing the production costs and a large diversity of languages available. However, from another perspective, it causes the reduction of

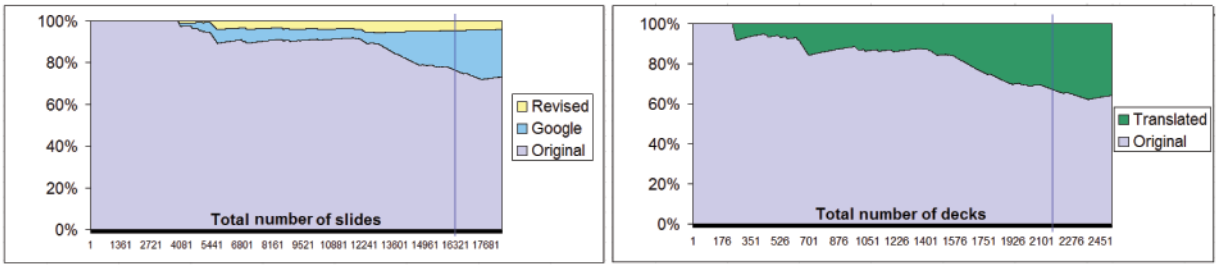

Figure 7: Percentage distribution between original and translated versions vs. total number of content objects for slides and decks. Blue line shows an example time moment, discussed in the paragraph above

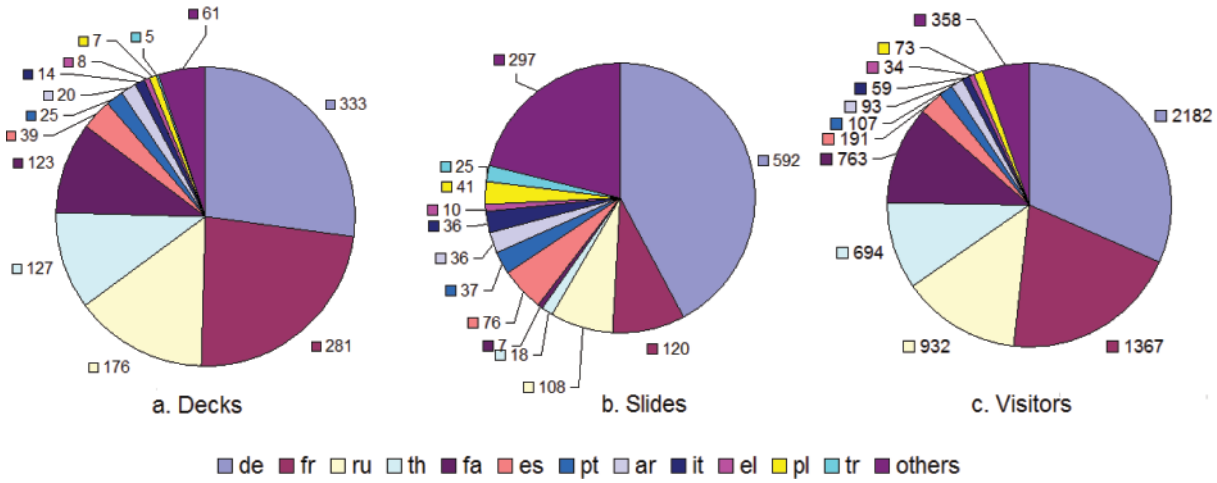

Figure 8: Distribution of languages for content objects and new visitors

average content quality, as refining the translation needs time and human resources. This is illustrated by the decreasing percentage of revised slides in comparison with translated ones. The solution is in additional user motivation to put effort in refining the concrete translations according to their knowledge in both domain and languages.

The diagrams on Figure 8 show the language distributions for decks, slides and visitors (according to Google analytics). Only unique visitors are counted. Due to its large percentage, we excluded English language from the resulting diagrams to increase readability. The statistics show the visible correlation between number of content objects available in a concrete language and number of visitors speaking the language (for the most of languages). The results look promising, as they prove the involvement of non- English-speakers into the global e-learning community activity. Especially promising is the fact that more than 13% of overall visitors belong to developing countries and regions (mostly Eastern Europe and Russia, Turkey, Arabic-speaking countries, Thailand). We believe that this percentage will increase with increasing popularity of the source.

Related work

Multilingual content management

Our work is based on previously done experiments and investigations. The most significant of these are summarized in O'Leary (2008). In particular, we based our work on the results of the experiment described in Nomura *et al. (*2003). This was one of the first studies on multilingual collaboration. The experiment was organized within a multilingual bulletin board system. It was found that machine translations were problematic, impairing communication. As a result, the authors propose to allow multilingual users to modify translated sentences to improve the overall level of the translation.

Modern research on multilingual content managing is mostly aimed to synchronize existing texts using NLP methods. For example, Monz *et al.* (2011) introduce a framework for content synchronization of wikis. However, all automated approaches of content synchronization face an important issue, namely, their application, especially in wikis, results in propagating incorrect statements from one language to all others. We consider this issue to be crucial in the educational domain, as it complicates the removal of outdated concepts from the content. In our semi-automatic approach of slide content synchronization, the incorrect sentences are not propagated between versions, because synchronization for each language has to be triggered and controlled manually. Therefore, outdated

concepts in one of the languages will not be accepted by the more advanced in the domain community members from another region.

Another approach, proposed in Negri *et al.* (2011), applies crowd-sourcing techniques to the creation of cross-lingual text entailment corpora. The paper is mainly aimed on improving the quality of translation and does not discuss the challenges of collaborative content authoring. However, we consider the study for future work, as we plan to increase the quality of translation, made by Google Translate Service.

Collaborative creation of e-learning content

The importance of creating reusable and re-purposable e-learning objects is widely accepted by the e-learning community (Devedzic, 2006). However, most of the works address the learning object reuse problem by means of semantic meta-data annotations, content tagging and packaging instead of by creating richly structured, reusable learning objects from the ground. The importance of creating learning objects with reuse already in mind was, for example, stated by Pedreira, Méndez and Martínez (2009, p. 532): "Content (...) should be represented not as an object of study but rather as necessary elements towards a series of objectives that will be discovered in the course of various tests." There are only a few approaches to the direct authoring of reusable content, such as, for example, learning examples creation (Kuo *et al.*, 2008) or semantic structuring and annotation of video fragments (Barriocanal *et al.*, 2011).

We should also mention two Learning Objects Repositories (LORs), that allow to produce structured reusable content. The first of them, Connexions (http://cnx.org/), presents the learning material as a combination of paragraphs, of which each could be easily edited or deleted. However, this structuring is done more for comfortable editing and does not have any functional benefits; the paragraphs cannot be reused or annotated independently. Thus, Connexions presents only an improved user interface for wiki-based systems. The second example of structuring, which is closer to our idea, is LeMill (http://lemill.net/). LeMill provides a way of collaborative editing of presentations by implementing presentations as a group of images which can be edited collaboratively. However, to edit a slide, a user has to replace it with another one. Also, it is impossible to have several subgroups of slides within a presentation. Searching through the slides (not presentations) is also not implemented. This means that slides can not be manipulated as independent learning objects.

The CrowdLearn concept differs from the existing approaches for managing e-learning content. It enables the construction of semantically structured learning objects from existing sources by combining, reordering and simple editing.

Wiki-based collaborative knowledge engineering

The importance of wikis for collaborative knowledge engineering is widely acknowledged. In (Richards, 2009), for example, a knowledge engineering approach which offers wiki-style collaboration is introduced, aiming to facilitate the capture of knowledge-in-action which spans both explicit and implicit knowledge types. The approach extends a combined rule and case-based knowledge acquisition technique, known as Multiple Classification Ripple Down Rules, to allow multiple users to collaboratively view, define and refine a knowledge base over time and space. In a more applied context, Haake, Lukosch and Schümmer (2005) introduce the concept of wiki templates that allow end-users to define the structure and appearance of a wiki page in order to facilitate the authoring of structured wiki pages. Similarly the hybrid wiki approach (Matthes, Neubert & Steinhoff, 2011) aims to solve the problem of using (semi-)structured data in wikis by means of page attributes. In our approach we apply the wiki paradigm to the creation and collaboration around (semi-)structured learning objects.

Conclusions

In this paper we presented CoSMEC, a concept for organizing collaborative authoring of multilingual educational resources. CoSMEC leverages the conversion of educational resources into multilingual content objects. Versions in different languages are inter-connected and can be semi-automatically synchronized. Our implementation proves the viability of the concept and shows possible directions for further work. For example, the Google Translate API does not allow users to create domain vocabularies or choose the best term like the Google Translate online service. This decreases the quality of the initial translation and can sometimes result in not-understandable content. Another field for future work is matching users with content they might help to translate. Such matching should take into account not only language knowledge, but domain experience and interests of the users as well. This can then motivate users to a higher degree to revise translations and improve the quality of the learning content.

Acknowledgment

This paper was presented at the 2014 OpenCourseWare Consortium Global Conference, held in Ljubljana (Slovenia) in April 23th-25th 2014 (http://conference.ocwconsortium.org/2014), with whom *Open Praxis* established a partnership. After a pre-selection by the Conference Programme Committee, the paper underwent the usual peer-review process in *Open Praxis*.

References

ADL (2011a). *SCORM 2004 4th Edition Specification. Technical report*. Retrieved from http://www.adlnet.gov/wp-content/uploads/2011/07/SCORM_2004_4ED_v1_1_Doc_Suite.zip

ADL (2011b). *SCORM Users guide for programmers. Technical report*. Retrieved from http://www.adlnet.gov/wp-content/uploads/2011/12/SCORM_Users_Guide_for_Programmers.pdf

Barriocanal, E. G., Sicilia, M. Á., Alonso, S. S. & Lytras, M. D. (2011). Semantic annotation of video fragments as learning objects. *Interactive Learning Environments*, *19*(1), 25–44. http://dx.doi.org/10.1080/10494820.2011.528879

Devedzic, V. (2006). *Semantic Web and Education*. Secaucus, NJ, USA: Springer. Retrieved from http://books.google.de/books?id=Rjdpb5wQu38C

Haake, A., Lukosch, S. & Schümmer, T. (2005). Wiki-templates: adding structure support to wikis on demand. *Proceedings of the 2005 International Symposium on Wikis*, 41–51. ACM. http://dx.doi.org/10.1145/1104973.1104978

Kuo, Y. H., Kinshuk, Q. T., Huang, Y. M., Liu, T. C. & Chang, M. (2008). Collaborative creation of authentic examples with location for u-learning. *E-Learning,* 16–20. Retrieved from http://adapt.athabascau.ca/publications/papers/el2008_kuo.pdf

Linta, S. R., Khan, R. & Ahmed, F. (2011). *Towards E-Learning Management System Using Semantic Web Technologies: A Great Proposed Model for E-LMS in Semantic Web and a Unique University Namespace "Univ" for Developing This E-Lms.* Germany: LAP Lambert Academic Publishing. Retrieved from http://dl.acm.org/citation.cfm?id=2132641

Matthes, F., Neubert, C. & Steinhoff, A. (2011). Hybrid wikis: Empowering users to collaboratively structure information. *ICSOFT*, 250–259. Retrieved from https://wwwmatthes.in.tum.de/file/1cpkdmgrqgul6/sebis%20Public%20Website/_/Ma11a%20-%20Hybrid%20Wikis%20-%20%20Empowering%20Users%20to%20Collaboratively%20Structure.../Ma11a.pdf

Monz, C., Nastase, V., Negri, M., Fahrni, A., Mehdad, Y. & Strube, M. (2011). Cosyne: a framework for multilingual content synchronization of wikis. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 217–218. http://dx.doi.org/10.1145/2038558.2038601

Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D. & Marchetti, A. (2011). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 670–679. Retrieved from http://www.mt-archive.info/EMNLP-2011-Negri.pdf

Nomura, S., Ishida, T., Yasuoka, M., Yamashita, N. & Funakoshi, K. (2003). Open source software development with your mother language: Intercultural collaboration experiment 2002. *HCI2003*, 1163–1167. Retrieved from http://www.ai.soc.i.kyoto-u.ac.jp/ice/pub/Econference1_HCI_final_20021030.pdf

O'Leary, D. E. (2008). A multilingual knowledge management system: A case study of FAO and WAICENT. *Decision Support Systems*, 45(3), 641–661. http://dx.doi.org/10.1016/j.dss.2007.07.007

Pedreira, N., Méndez, J. R. & Martínez, M. (2009). E-learning in new technologies. J. R. Rabuñal *et al.* (eds.). *Encyclopedia of Artificial Intelligence* (pp. 532–535). IGI-Global. http://dx.doi.org/10.4018/978-1-59904-849-9.ch081

Richards, D. (2009). A social software/web 2.0 approach to collaborative knowledge engineering. *Information Sciences, 179*(15), 2515–2523. http://dx.doi.org/10.1016/j.ins.2009.01.031

Tarasowa, D., Khalili, A., Auer, S. & Unbehauen, J. (2013). Crowdlearn: Crowd-sourcing the creation of highly-structured e-learning content. *Proceedings of the International Conference on Computer Supported Education*, 33–42. http://dx.doi.org/10.5220/0004384100330042